Opinion and Polarity Detection within Far-East Languages in NTCIR-7

Olena ZUBARYEVA Jacques SAVOY
Computer Science Dept, University of Neuchatel
rue Emile Argand 11, 2009 Neuchatel, Switzerland
{ Olena.Zubaryeva, Jacques.Savoy }@unine.ch

Abstract

This paper presents our work in the Multilingual Opinion Analysis Task (MOAT) done during the NTCIR-7 workshop. This is our first participation in this kind of retrieval and classification task in which we participated for the English, Japanese and traditional Chinese language. As a basic model we suggested a probabilistic model derived from Muller's method [1] that allows us to determine and weight terms (isolated words, bigram of words, noun phrases, etc.) belonging to a given category compared to the rest of the corpus. In the current task, the classification categories are positive, negative, neutral and not opinionated. To succeed at this classification task, we have adopted the logistic regression method in order to define the most probable category for each input sentence. Our participation was strongly motivated by the objective to suggest an approach on the polarity subtask of the MOAT with a minimal linguistic component. **Keywords:** Opinion detection, polarity classification, classification model, word distribution, opinionated IR, logistic regression.

1 Introduction

With the broad specter of available information on the web and the growth of user-focused activity of adding content on various subjects, the task of detecting opinions and polarity of those opinions has raised interest in the research community. It is of high interest to develop a system that would be adaptable for different languages to detect opinionated documents on the one hand, and on the other to be able to detect their polarity (positive, negative or neutral). This task is important in many areas of Natural Language Processing (NLP) [2] from question/answering (Q/A), document summarization, especially with the increasing potential application in several web-oriented domains.

The NTCIR-7 MOAT (Multilingual Opinion Analysis Task) defined sentences as the information items. The four subtasks included the opinion detection, if the opinion was detected its opinion holder/holders and target, relevance and polarity. The required subtask was to identify the presence of opinion in the sentence. All other subtasks were optional. This is our first participation and we have participated in all subtasks except the identification of the opinion holder and opinion target. We consider these latter subtasks to be more challenging and requiring more sophisticated NLP tools depending more heavily on the underlying natural language.

Our main goal in the NTCIR-7 Multilingual Opinion Analysis Task was the first approbation of our system on effective retrieval of opinionated information in different languages. We want to promote an effective search system in which the linguistic component could be both clearly identified. Thus to achieve this goal, we have participated in the English, traditional Chinese and Japanese language tracks.

The remainder of the paper is organized in the following way. Section 2 presents related work while Section 3 exposes our approach to determine the opinion and polarity of the sentences. We present the results and their evaluation in Section 4. Finally, the conclusions and future work are given in Section 5.

2 Related Work

The main objective of our participation in the MOAT task is to develop automatic retrieval and classification scheme that will be able to first to retrieve short information items (e.g., sentences, short paragraphs) according to a submitted query. In the second stage, the system must classify them according to their opinionated content as factual (no opinion), positive, negative and neutral (presenting mixed opinions). The focus in our participation in the NTCIR-7 was to propose a general approach that can be easily deployed for different natural languages.

We must first recognize that classifying short information items into positive, negative and neutral opinion categories is a difficult task, due to the fact that the semantic differences between the category neutral and the two others could be small leading to complex problems when designing and implementing an effective discrimination function. Moreover, the distinction between positive or negative could be denoted by a small element in the underlying text (e.g., a simple "not"). Finally, the distinction between neutral and either positive or negative could sometimes be questionable for a human being, as well as evaluating whether or not a given sentence (or short paragraph) conveys an opinion is not.

When viewing an opinion-finding task as a classification task (after retrieving the relevant items), it is usually considered a supervised learning problem where a statistical model performs a learning task by analyzing a pool of labeled documents. Two questions must be solved, namely defining an effective classification algorithm [3] and determining pertinent features that might effectively discriminate between opinionated and factual sentences / paragraphs.

From this perspective, during the two last TREC opinion-finding tasks [4], [5] and last NTCIR workshop [6], a series of suggestions surfaced. Based on the English grammar, Levin defined different verb categories (characterize, declare, conjecture, admire, judge, assess, say, complain, advise) and their features (a verb corresponding to a given category occurring in the analyzed information item) that may be pertinent as a classification feature [7]. However, words such as these cannot always work correctly as clues, for example with the word "said" in the two sentences "The iPhone price is expensive, said Ann" and "The iPhone price is 600 $, said Ann." Both sentences contain the clue word "said" but only the first one contains an opinion on the target product.

We might also mention OpinionFinder [8], a more complex system that performs subjectivity analyses to identify opinions as well as sentiments and other private states (speculations, dreams, etc.). This system is based on various classical computational linguistics components (tokenization, part-of-speech (POS) tagging [9], [10] as well as classification tools. For example, a naive Bayes classifier [3] is used to distinguish between subjective and objective sentences. A rule-based system is included to identify both speech events ("said," "according to") and direct subjective expressions ("is happy," "fears") within a given sentence. Of course such learning system requires both a training set and a deeper knowledge of a given natural language (morphological components, syntactic analyses, semantic thesauri).

The lack of enough training data for all these learning-based sub-systems is clearly a drawback, although not all groups participating in the pilot NTCIR-6 opinion analysis task encountered this same problem. Moreover, it is difficult to objectively establish when a complex learning system has enough training data (and to objectively measure the amount of training data needed in a complex ML model).

3 Our Opinion-Detection Approach

Our system is based on two components, namely the extraction of useful features (isolated words in this study) to allow an effective classification, and second a classification scheme [3]. Our system uses word forms (tokens) to perform sentence identification within the two classes. As shown by Kilgarriff [11], the selection of words (or in general features) in an effort to characterize a particular category is a difficult task, when analyzing and criticizing various statistical measures [12], [13], [14]. The selection and weighting of words is explained in Section 3.1 while Section 3.2 exposes the main aspects of our classification scheme based on logistic regression [15].

3.1 Features Extraction

In order to determine the features that can help distinguishing between factual and opinionated documents in one hand, and on the other between the polarities of the sentences, we have selected the tokens. The goal is therefore to design a method capable of selecting terms that clearly belong to one type of polarity compared to the other possibilities. Various authors have suggested formulas that could meet this objective under the condition that we use words and their frequencies or distributions [12], [13], [11], [14]. These suggested approaches are usually based on a contingency table (see Table 1).

	S	C-	
ω	*a*	*b*	*a*+*b*
not ω	*c*	*d*	*c*+*d*
	a+*c*	*b*+*d*	*n*=*a*+*b*+*c*+*d*

Table 1. Example of a contingency table.

In this table, the letter *a* represents the number of occurrences (tokens) of the word ω in the document set S (corresponding to a subset of the larger corpus C). The letter *b* denotes the number of tokens of the same word ω in the rest of the corpus (denoted C-) while *a*+*b* is the total number of occurrences in the entire corpus (denoted C). Similarly, *a*+*c* indicates the total number of tokens in S. The entire corpus C corresponds to the union of the subset S and C- (C = S∪C-) that contains *n* tokens (*n* = *a*+*b*+*c*+*d*).

Based on the MLE (Maximum Likelihood Estimation) principle the values shown in a contingency table could be used to estimate various probabilities. For example we might calculate the probability of the occurrence of the word ω in the entire corpus C as Pr(ω) = (*a*+*b*)/*n* or the probability of finding in C a word belonging to the set S as Pr(S) = (*a*+*c*)/*n*.

Now to define the discrimination power a term ω, we suggest deriving a weight attached to it according to Muller's method [1]. We assume that the distribution of the number of tokens of the word ω follows a binomial distribution [16] with the parameters *p* and *n'*. The parameter *p* represented the probability of drawing the word ω (or Pr(ω)) and could be estimated as (*a*+*b*)/*n*. If we repeat this drawing *n'* = *a*+*c* times, we will have an estimate of the number of word ω included in the subset S as Pr(ω)·*n'*. On the other hand, Table 1 gives also the number of observations of the word ω in S, and this value is denoted by *a*. A large difference between *a* and the product Pr(ω)·*n'* is clearly an indication that the presence of *a* occurrences of the term ω is not due by chance but corresponds to an intrinsic characteristic of the set S compared to the set C-.

In order to obtain a clear rule, we suggest computing the Z score attached to each word ω. If the mean of a binomial distribution is Pr(ω)·*n'*, its variance is *n'*·Pr(ω)·(1-Pr(ω)). These two elements are needed to compute the standard score as described in Equation 1.

$$Zscore(\omega) = \frac{a - n`\cdot \Pr(\omega)}{\sqrt{n`\cdot \Pr(\omega)\cdot(1-\Pr(\omega))}} \qquad (1)$$

Using the MOAT-NTCIR 6 English corpus as an example (and as the training data), Table 2 indicates that the word "said" occurs 561 in opinionated sentences and 241 in the rest of the corpus composed of factual sentences (for a total of 802 tokens). The opinionated part contains 69,885 tokens, representing around 55.8% of the total number (125,226 tokens). Clearly, we encountered more often the word "said" in the opinionated sentences (561 times) that the simple proportion (441 = 55% of 802). The Z score for this term is equal to 5.34, indicating clearly an overuse of this term in the opinionated sentences.

	opinionated	rest	
"said"	561	241	802
- "said"	69,324	55,100	124,424
	69,885	55,341	125,226

Table 2. Example with the word "said" in the opinionated and the whole English corpus

As a decision rule we consider the words having a Z score between -2 and 2 as terms belonging to a common vocabulary, as compared to the reference corpus (as for example "will," "with," "many," "friend," or "forced" in our example). This threshold was chosen arbitrary. A word having a Z score > 2 would be considered as overused (e.g., "that," "should," "must," "not," or "government" in MOAT-NTCIR 6 English corpus), while a Z score < -2 would be interpreted as an underused term (e.g., "police," "cell," "year," "died," or "according"). The arbitrary threshold limit of 2 corresponds to the limit of the standard normal distribution, allowing us to only find 5% of the observations (around 2.5% less than -2 and 2.5% greater than 2). As shown in Figure 1, the difference between our arbitrary limit of 2 (drawn in solid line) and the limits delimiting the 2.5% of the observations (dotted line) are rather close.

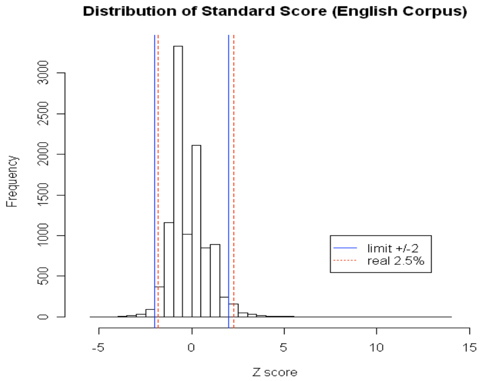

Figure 1. Distribution of the Z score (MOAT-NTCIR 6 English corpus, opinionated)

Based on a training sample, we were able to compute the Z score for different words and retain only those having a large or small Z score value. Such a procedure is repeated for all classification categories (e.g., positive, negative and neutral in the current context). It is worth mentioning that such a general scheme may work with isolated words (as applied here) or *n*-grams (that could be a sequence of either characters or words), punctuations or other symbols (numbers, dollar signs), syntactic patterns (e.g., verb-adjective) or other features (presence of proper names, hyperlinks, etc.)

3.2 Our Classification Model

When our system needs to determine the polarity of a sentence, we first represent this sentence as a set of words. For each word, we can then retrieve the Z scores for each category. If all Z scores for all words are judged as belonging to the general vocabulary, our classification procedure selects the default category. If not, we may increase the weight associated with the corresponding category (e.g., for the positive class if the underlying term is overused in this category).

Such a simple additive process could be viewed as a first classification scheme, selecting the class having the highest score after enumerating all words occurring in a sentence. For this model, we can define three variables, namely *SumPos* indicating the sum of the Z score of terms overused in positive class (i.e. Z score > 2) and appearing in the input sentence. Similarly, we can define *SumNeg*, and *SumNeutral* for the other two classes. As additional explanatory variables, we also use the 8 characteristic term statistics to calculate the corresponding polarity score for each sentence. The scores are calculated by applying the following formulae:

$$Pos\_score = \frac{\#PosOver}{\#PosOver + \#PosUnder}$$
$$Neg\_score = \frac{\#NegOver}{\#NegOver + \#NegUnder} \qquad (2)$$
$$Neutral\_score = \frac{\#NeuOver}{\#NeuOver + \#NeuUnder}$$

in which *#PosOver* indicated the number of terms in the evaluated sentence that tends to be overused in positive documents (i.e. Z score > 2) while *#PosUnder* indicated the number of terms that tend to be underused in the class of positive documents (i.e. Z score < -2). Similarly, we can define the variables *#NegOver, #NegUnder, #NeuOver, #NeuUnder,* but for their respective categories, namely negative and neutral. The score is defined as the logistic transformation π(x) given by each logistic regression model defined as:

$$\pi(x) = \frac{e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{12} \beta_i x_i}} \qquad (3)$$

where β_i are the coefficients obtained from the fitting and x_i are the variables, and k is the number of variables. These coefficients reflect the relative importance of each explanatory variable in the final score.

For each sentence, we can compute the π(x) corresponding to the three possible categories and the final decision is simply to classify the sentence according to the max π(x) value. This approach takes account of the fact that some explanatory variables may have more importance than other in assigning the correct category. However, we must recognize that the length of the underlying sentence is not directly taken into account in this first model. Our underling assumption is that all sentences have a similar number of indexing tokens.

4 Evaluation

In order to evaluate the capability of an automatic system to retrieve and classify correctly different information items, we may impose that the answers are a ranked list and then evaluate the system's performance according to classical IR measures such as MAP. This approach was adopted during the last Blog tracks at TREC [4], [5]. As another approach we may evaluate the classification performance based on a set-based approach, judging the system's capability to identify the different categories. The traditional evaluation measures based on sets (precision, recall, F-measure) can then be applied. This choice was made for the NTCIR workshops [6] and explained in the current workshop [17].

On the other hand, we have assumed until now that words can be extracted from a sentence in order to define the needed features used to determine if the underlying information item conveys an opinion or not. Working with the Japanese or Chinese languages this assumption does no longer hold and we need to determine indexing units by either applying an automating segmentation approach (based either on a morphological (e.g., CSeg&Tag) or a statistical method [18] or considering *n*-gram indexing approach (unigram, bigram or both unigram and bigram). Finally we may also consider a combination of both *n*-gram and word-based indexing strategies [10], [18].

4.1 Traditional Chinese Language

We participated in the traditional Chinese language task and were able to submit one run based on our first classification model. Based on our past IR experiments [19], we have selected a combined unigram & bigram indexing scheme for this language.

		Prec.	Recall	F-mes
Relevance	Lenient	0.961	0.846	0.900
	Strict	0.875	0.844	0.859
Opinion	Lenient	0.543	0.927	0.685
	Strict	0.692	0.938	0.797
Polarity	Lenient	0.233	0.398	0.294
	Strict	0.307	0.416	0.353

Table 3: MOAT evaluation for the traditional Chinese opinion analysis

4.2 Japanese Language

With the Japanese language we submitted a single run based, as for the Chinese language, on our first classification model. Based on our past experiment [19], we have selected a bigram indexing scheme for this language.

		Prec.	Recall	F-mes
Relevance	Lenient	0.415	0.192	0.262
	Strict	0.155	0.146	0.151
Opinion	Lenient	0.536	0.200	0.291
	Strict	0.416	0.213	0.281
Polarity	Lenient	0.325	0.055	0.094
	Strict	0.291	0.050	0.085

Table 4: MOAT evaluation for the Japanese opinion analysis

4.3 English Language

For the evaluation of sentences in English, the assumption of isolated words (bag-of-words) was used by our system. We were able to send three runs for this language; the third is based on the same classification model used for both the Chinese and Japanese languages. The second model is based on the extended logistic model that includes more explanatory variables. Specifically, we experimented with the logarithms of the initial variables on the training set. The first run used features of the second run and an additional query expansion approach used to better determine opinionated sentences about the target entity. Namely, we used around 500 words that identify speech events ("explained", "commented", etc.) or subjective expressions ("sympathized", "accused", etc.). Thus, using this query expansion technique we tried to identify sentences relevant to the query and possibly opinionated. For this set of sentences that were not classified as opinionated by our initial model, we judged them as opinionated with the polarity that has the highest score for the sentence.

	Lenient/Strict		Prec.	Recall	F-mes
Relevance	Model 1	L	0.417	0.599	0.492
		S	0.161	0.677	0.261
	Model 2	L	0.342	0.454	0.390
		S	0.143	0.563	0.228
	Model 3	L	0.331	0.433	0.375
		S	0.138	0.537	0.220
Opinion	Model 1	L	0.332	0.700	0.450
		S	0.105	0.743	0.184
	Model 2	L	0.377	0.576	0.456
		S	0.120	0.613	0.200
	Model 3	L	0.383	0.553	0.453
		S	0.123	0.594	0.203
Polarity	Model 1	L	0.228	0.367	0.281
		S	0.064	0.417	0.111
	Model 2	L	0.246	0.319	0.278
		S	0.069	0.360	0.115
	Model 3	L	0.250	0.310	0.277
		S	0.067	0.337	0.112

Table 5: MOAT evaluation for the three models used with the English corpus

As one can see form the results the first model that used query expansion technique as expected gave overall better performance in the relevance, opinion and polarity subtasks. This tendency suggests that probably more experiments with syntax and content identification heuristics should be used to improve the performance of the base statistical model.

5. Future work

In our system, we have suggested using a statistical method (Z score) to identify those terms that adequately characterize subsets of the corpus belonging to positive, negative, neutral or non-opinionated subsets. In this selection, we focused only on the statistical aspect (distribution difference) of words or bigrams. We also have demonstrated on the English subtask how we can use the query expansion to identify the possibility of opinion expressed in the sentences that otherwise were identified as not opinionated by the system.

This study was limited to single words but in further research we could easily consider longer word sequences to include phrases (both bigrams or trigrams as well as phrases identified by a POS tagger). We may also consider punctuations (e.g., quotation marks (""), question marks (?), exclamation points (!), etc.) as well as other symbols (e.g. $, mm, mainly associated with facts) to distinguish between factual and opinionated documents. The most useful terms would also then be added to the query to improve the rank of opinionated documents. As another approach, we could use the evaluation of co-occurrence terms of pronouns "I" and "you" mainly with verbs (e.g., "believe," "feel," "think," "hate") in order to boost the rank of retrieved items.

Using freely available POS taggers[1], we could take POS information into account [9], [10] and hopefully develop a better classifier. For example, the presence /occurrence of proper names and their frequency or distribution might help us classify a document as being opinionated or not. The presence /occurrence of adjectives and adverbs, together with their superlative (e.g., best, most) or comparative (e.g., greater, more) forms could also be useful hints regarding the presence of opinionated versus factual information.

6. Conclusion

For our first participation in a classification task, we have suggested a general method to define and weight isolated words in order to build a set of useful features able to classify sentences into different categories. Our classification scheme is based on the logistic regression method [15]. In our objective to propose a general classification scheme able to work with different natural languages, we have adapted our system to work with the English, Japanese and traditional Chinese languages.

The evaluation results obtained by our system are in the average and we need to analyze the results query-by-query to determine the most important reasons explaining the poor performance of our approach for some queries. On the other hand, we have other possibilities to be included in our classification scheme (e.g., bigram of words, noun phrase, punctuation, word categories) that could improve the efficiency of the suggested model. In this selection of good discrimination features, we have also to balance between purely

[1] For the Japanese language we can consider the MeCab software (see mecab.sourceforge.net), an advanced version of the ChaSen analyzer. We can combine this with the KAKASI system (see kakasi.namazu.org) or the Juman (nlp.kuee.kyoto-u.ac.jp/nl-resource) morphological analyzer.

statistical features (e.g., the letters distribution) having no direct interpretation and linguistic-based features that could be either difficult to find in a short sentence or that have no discrimination power beyond toy-examples.

ACKNOWLEDGMENTS

The authors would like to thank the NTCIR-7 task organizers for their efforts in developing test-corpus in opinionated IR. This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

References

[1] Muller, C. (1992). *Principes et méthodes de statistique lexicale*. Champion, Paris.

[2] Nugues, P.M. 2006. *An Introduction to Language Processing with Perl and Prolog*. Springer-Verlag, Berlin.

[3] Witten, I.A., & Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd Ed., Morgan Kaufmann, San Francisco (CA).

[4] Ounis, I., de Rijke, M., Macdonald, C., Mishne, G., & Soboroff, I. 2007. Overview of the TREC-2006 blog track. *Proceedings TREC-2006*, NIST Publication #500-272, 17-32.

[5] Macdonald, C., Ounis, I., & Soboroff, I. 2008. Overview of the TREC-2007 blog track. *Proceedings TREC-2007*, NIST Publication #500-274, 1-13.

[6] Seki, Y., Evans, D.K., Ku, L.W., Chen, H.H., Kando, N., & Lin, C.Y. 2007. Overview of opinion analysis pilot task at NTCIR-6. *Proceedings NTCIR-6*, National Institute of Informatics, 265-278.

[7] Bloom, K., Stein, S., & Argamon, S. 2007. Appraisal extraction for news opinion analysis at NTCIR-6. *Proceedings NTCIR-6*, National Institute of Informatics, 279-289.

[8] Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S., 2005. OpinionFinder: A system for subjectivity analysis. *Proceedings HLT/EMNLP*, Vancouver (BC), 34-35.

[9] Marcus, M.P., Santorini, B., & Marcinkiewicz, M.A. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313-330.

[10] Toutanova, K., & Manning, C. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagging. *Proceedings EMNLP / VLC-2000*, 63-70.

[11] Kilgarriff, A. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97-133.

[12] Church, K.W., & Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22-29.

[13] Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61-74.

[14] Manning, C.D., & Schütze, H. 2000. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

[15] Hosmer D., & Leneshow S., *Applied Logistic Regression*. Wiley Interscience, New York, 2000.

[16] Baayen, H.R. 2001. *Word Frequency Distributions*. Kluwer Academic Press, Dordrecht, NL.

[17] Seki, Y., Evans, D.K., Ku, L.W., Chen, H.H., Kando, N., & Lin, C.Y. 2008. Overview of opinion analysis pilot task at NTCIR-7. *Proceedings NTCIR-7*, National Institute of Informatics.

[18] Murata, M., Ma, Q., & Isahara, H. 2003. Applying multiple characteristics and techniques to obtain high levels of performance in information retrieval. *Proceedings of NTCIR Workshop3*.

[19] Savoy, J. 2005. Comparative study of monolingual and multilingual search models for use with Asian languages. *ACM Transactions on Asian Languages Information Processing*, 4:163-189.